https://doi.org/10.5664/jcsm.9174

JCSM | Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Interrater sleep stage scoring reliability between manual scoring from two European sleep centers and automatic scoring performed by the artificial intelligence–based Stanford-STAGES algorithm

Matteo Cesari, PhD[1]; Ambra Stefani, MD[1]; Thomas Penzel, PhD[2,3]; Abubaker Ibrahim, MD[1]; Heinz Hackner[1]; Anna Heidbreder, MD[1]; András Szentkirályi, MD, PhD[4]; Beate Stubbe, MD[5]; Henry Völzke, MD, PhD[6]; Klaus Berger, MD[4]; Birgit Högl, MD[1]

[1]*Department of Neurology, Medical University of Innsbruck, Innsbruck, Austria;* [2]*Interdisciplinary Sleep Medicine Center, Charité-Universitätsmedizin Berlin, Berlin, Germany;* [3]*Saratov State University, Saratov, Russian Federation;* [4]*Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany;* [5]*Department of Internal Medicine B, University Medicine Greifswald, Greifswald, Germany;* [6]*Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany*

Study Objectives: The objective of this study was to evaluate interrater reliability between manual sleep stage scoring performed in 2 European sleep centers and automatic sleep stage scoring performed by the previously validated artificial intelligence–based Stanford-STAGES algorithm.

Methods: Full night polysomnographies of 1,066 participants were included. Sleep stages were manually scored in Berlin and Innsbruck sleep centers and automatically scored with the Stanford-STAGES algorithm. For each participant, we compared (1) Innsbruck to Berlin scorings (INN vs BER); (2) Innsbruck to automatic scorings (INN vs AUTO); (3) Berlin to automatic scorings (BER vs AUTO); (4) epochs where scorers from Innsbruck and Berlin had consensus to automatic scoring (CONS vs AUTO); and (5) both Innsbruck and Berlin manual scorings (MAN) to the automatic ones (MAN vs AUTO). Interrater reliability was evaluated with several measures, including overall and sleep stage-specific Cohen's κ.

Results: Overall agreement across participants was substantial for INN vs BER ($\kappa = 0.66 \pm 0.13$), INN vs AUTO ($\kappa = 0.68 \pm 0.14$), CONS vs AUTO ($\kappa = 0.73 \pm 0.14$), and MAN vs AUTO ($\kappa = 0.61 \pm 0.14$), and moderate for BER vs AUTO ($\kappa = 0.55 \pm 0.15$). Human scorers had the highest disagreement for N1 sleep ($\kappa_{N1} = 0.40 \pm 0.16$ for INN vs BER). Automatic scoring had lowest agreement with manual scorings for N1 and N3 sleep ($\kappa_{N1} = 0.25 \pm 0.14$ and $\kappa_{N3} = 0.42 \pm 0.32$ for MAN vs AUTO).

Conclusions: Interrater reliability for sleep stage scoring between human scorers was in line with previous findings, and the algorithm achieved an overall substantial agreement with manual scoring. In this cohort, the Stanford-STAGES algorithm showed similar performances to the ones achieved in the original study, suggesting that it is generalizable to new cohorts. Before its integration in clinical practice, future independent studies should further evaluate it in other cohorts.

Keywords: automatic scoring; deep neural networks; computerized analysis; interrater variability; study of health in Pomerania; slow wave activity

Citation: Cesari M, Stefani A, Penzel T, et al. Interrater sleep stage scoring reliability between manual scoring from two European sleep centers and automatic scoring performed by the artificial intelligence–based Stanford-STAGES algorithm. *J Clin Sleep Med*. 2021;17(6):1237–1247.

BRIEF SUMMARY

Current Knowledge/Study Rationale: Recent years have seen an increasing number of automatic methods based on artificial intelligence to perform sleep stage scoring. Generalizability to new datasets constitutes a key factor for integration of these algorithms in clinical practice. Here, we independently investigate generalizability of the previously validated Stanford-STAGES algorithm to a new cohort of more than 1,000 participants.

Study Impact: To our knowledge, this is the first study proposing a large-scale and independent evaluation of a previously validated algorithm for sleep stage scoring. The Stanford-STAGES algorithm seems to be generalizable to new unseen cohorts, but other studies should further validate it in new cohorts with different patient groups, with the perspective of integrating it in clinical practice.

INTRODUCTION

According to international standards redacted by the American Academy of Sleep Medicine (AASM),[1] sleep evaluation is performed by recording of video polysomnography (PSG) and sleep epochs lasting 30 seconds are manually scored by human sleep experts as either wakefulness (W), rapid eye movement (REM) sleep, non–REM sleep stage 1 (N1) sleep, non–REM sleep stage 2 (N2) sleep, or non–REM sleep stage 3 (N3) sleep. Manual sleep stage scoring has some important drawbacks, including time consumption and the requirement of highly trained personnel. Furthermore, as AASM rules to score sleep are prone to a degree of subjective interpretation, sleep staging is prone to interrater variability.[2–7]

Automatic sleep staging methods would significantly reduce the time needed for analysis of video PSGs and would also overcome the problem of interrater variability.[8] Starting from the late 1960s,[9] hundreds of different methods for automated sleep stage scoring have been proposed (see Penzel et al,[10] Boostani et al,[11] Lajnef et al,[12] and Fiorillo et al[13] for comprehensive

reviews). However, such methods are still not used in clinical routine. This is because of many reasons, which include (1) the fact that most methods were validated only in cohorts with young healthy controls and are thus unreliable when applied to patients with sleep disorders; and (2) the lack of validation of these methods in different cohorts, thus not ensuring their generalizability to different populations.[13,14]

Because of the increase of computational power and the availability of large datasets, the last years saw the increasing development of automatic methods that use artificial intelligence techniques (ie, based on deep neural networks). Compared with automated methods based on other techniques, algorithms based on artificial intelligence can easily deal with large amount of data and can directly learn patterns from data, without the need of humans to define relevant features for correct sleep stage classification. These methods have shown promising performances in terms of agreement with manual sleep stage scoring.[13]

Among these methods, the Stanford-STAGES algorithm[15] (a software program written in Python performing both automated sleep stage scoring as well as automatic identification of narcolepsy patients based on automatic analysis of PSG signals; available at https://github.com/stanford-stages/stanford-stages, accessed March 5, 2020) seems to be a promising tool to be introduced in clinical sleep practice,[16] mainly for 2 reasons. First, it showed good agreement with manual sleep stage scoring in 4 different databases including healthy controls and patients.[15] Second, the algorithm describes sleep as a dynamic process, where each sleep epoch is represented as a mixture of W, N1, N2, N3, and REM sleep. Such a dynamic way of representing sleep has been shown to hold relevant physiologic information to identify patients with narcolepsy type 1[15] and might also help to identify and diagnose other sleep disorders.

However, the Stanford-STAGES algorithm has only been tested in 1 study, and thus does not guarantee its generalizability to other cohorts and populations. The aim of this study is to evaluate interrater reliability (IRR) for sleep stage scoring between human scorers from 2 European sleep centers and the automatic Stanford-STAGES algorithm in a cohort of more than 1,000 participants.

METHODS

Study participants

The participants included in this study were part of Study of Health in Pomerania-TREND, which is 1 of the 2 cohorts within the framework of the Study of Health in Pomerania[17] in northeastern Germany. Of the 4,420 participants included in the Study of Health in Pomerania-TREND baseline examinations, 1,249 underwent an optional 1-night full PSG. Because 183 participants were excluded for technical issues, our final study population included 1,066 participants. The population-based study was approved by the ethical committee of the University of Greifswald, Germany. The presented analysis was additionally approved by the ethical committee of the Medical University of Innsbruck, Austria.

PSG recordings and manual sleep scoring

Specific details concerning PSG recordings have been previously described.[18] Briefly, PSGs were recorded according to the AASM 2007 standards[19] with ALICE 5 devices (Philips Respironics, Eindhoven, The Netherlands), and the recordings included electroencephalogram (EEG; 6 derivations: F4A1, C4A1, O2A1, F3A2, C3A2, and O1A2), 2 electrooculographic channels, electromyogram recorded at chin and both anterior tibialis muscles, 1 electrocardiogram channel, pulse oximetry, nasal pressure, inductive plethysmography detecting respiratory effort, tracheal microphone, and body position sensor. EEG, electrooculographic, and electromyogram signals were sampled at 200 Hz. PSG recordings were digitally transferred in European Data Format (EDF) to the University Hospital Charité, Center of Sleep Medicine Berlin, Germany, where 30-second sleep epochs were manually scored by certified sleep technicians according to the AASM 2007 criteria.[18,19] Scoring of respiratory events was also performed according to AASM 2007,[18,19] and the apnea-hypopnea index (AHI) was calculated for each participant. As part of a research collaboration, the EDF files were also transferred to the Sleep Laboratory of the Department of Neurology, Medical University of Innsbruck (Austria), where the same sleep 30-second epochs were manually scored by 1 experienced sleep technician according to AASM 2012 criteria.[20,21] Furthermore, limb and periodic limb movements (PLMS) were scored and the PLMS index calculated for each participant.[21]

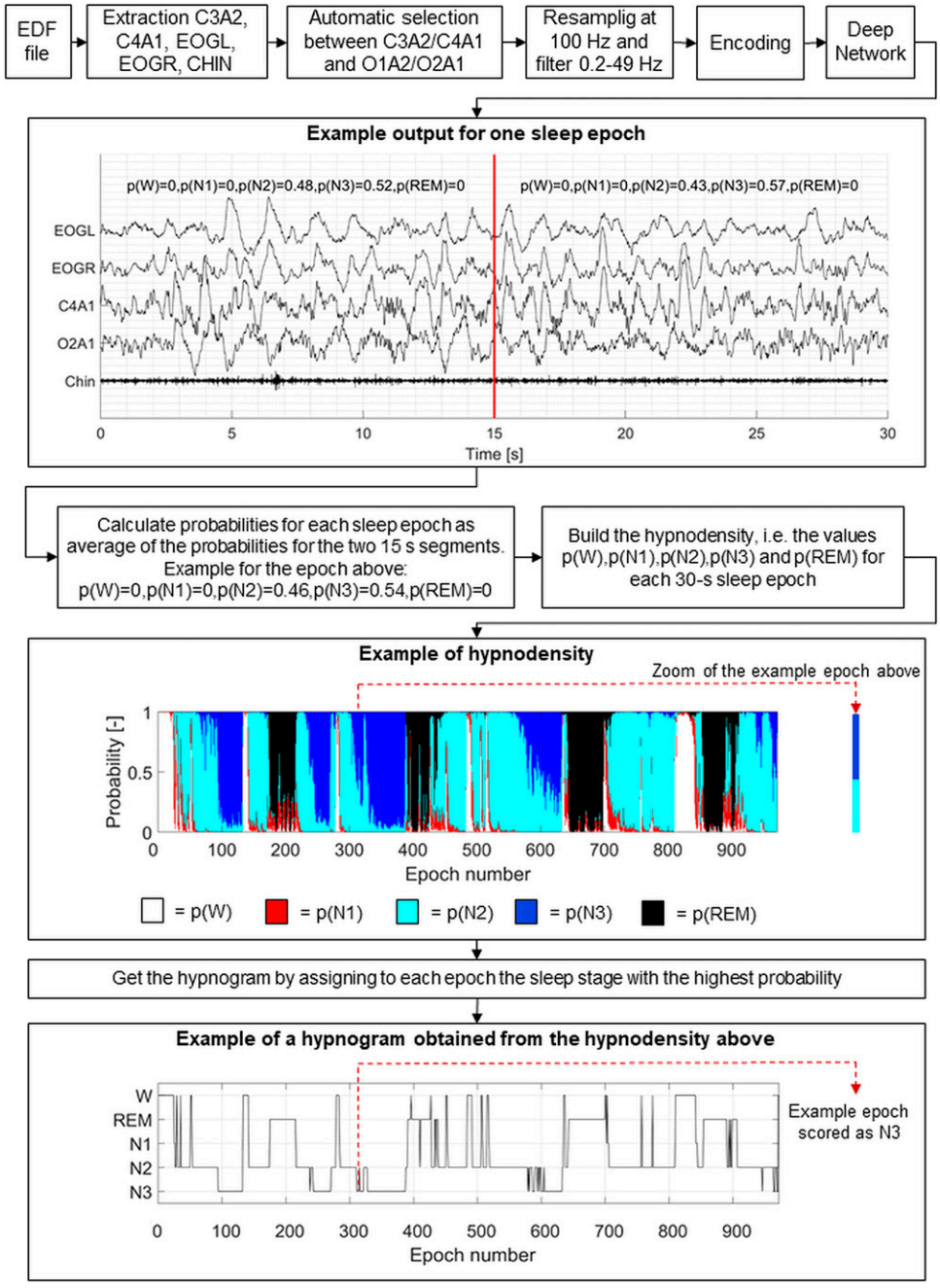
Automatic sleep stage scoring

To automatically score PSGs, the Python code of the Stanford-STAGES algorithm was retrieved from the repositories indicated in the original publication[15] (https://github.com/stanford-stages/stanford-stages, accessed March 5, 2020), and the provided instructions for installation and setup of the code were carefully followed to ensure correct implementation.

Figure 1 shows an overview of the steps of the automatic scoring. EDF files were given as input to the algorithm, which automatically performed the following steps. First, chin electromyogram, left and right electrooculographic signals, and 1 central (C3 or C4) and 1 occipital (O1 or O2) EEG signals referencing to the contralateral mastoid (A2 or A1, respectively) were extracted from the EDF file. The algorithm automatically selected whether to use the left or right EEG channels by choosing the less noisy ones following the procedure described in the original work.[15] Afterward, the signals were resampled at 100 Hz, band-pass filtered between 0.2 and 49 Hz, and encoded with cross-correlation. The encoded signals were given as input to 16 already trained long short-term memory deep networks. For each 15-second segment of the EDF file, the algorithm returned the probabilities that that segment was W, N1, N2, N3, and REM sleep.

To allow comparison of the automatic sleep stage scorings to the manual ones, we averaged the probabilities of W, N1, N2, N3, and REM sleep over the two 15-second segments included in a 30-second epoch. The values of sleep stages probabilities for each epoch allowed us to generate a

Figure 1—Schematic overview of the automatic sleep stage scoring with the Stanford-STAGES algorithm.



From the EDF files, the C3A2, C4A1, O2A1, and O1A2 electroencephalographic channels were extracted, as well as the electromyographic chin channel and the left and right electrooculographic channels. The algorithm automatically selected which of the 2 central and occipital channels to use. Then, the signals were resampled at 100 Hz, filtered between 0.2 and 49 Hz, and encoded with cross-correlation. The encoded signals were given in input to the deep neural network. For each 15-second segment, the network returned the probabilities that such segment was wakefulness (p(W)), N1 sleep (p(N1)), N2 sleep (p(N2)), N3 sleep (p(N3)), and rapid eye movement sleep (p(REM)). The figure reports an example epoch for which the obtained probability values are shown. For each 30-second sleep epoch, the average values of probabilities across the two 15-second segments were calculated, thus obtaining the values of probabilities for each sleep epoch. The hypnodensity was obtained as the graphical representation of the sleep stage probabilities for each sleep epoch. From the hypnodensity, the hypnogram was built by scoring each sleep epoch as the sleep stage with the highest probability. EDF = European data format; EOGL = electrooculogram left; EOGR = electrooculogram right; N1 = non-REM stage 1 sleep; N2 = non-REM stage 2 sleep; N3 = non-REM stage 3 sleep; REM = rapid eye movement sleep; W = wakefulness.

hypnodensity graph. The hypnogram was obtained by assigning to each epoch the sleep stage with the highest probability. **Figure 2, C and D**, shows the automatic hypnogram and hypnodensity obtained for 1 participant in comparison to the manual hypnograms scored in Innsbruck (**Figure 2A**) and in Berlin (**Figure 2B**).

Further details concerning the methodology of the Stanford-STAGES algorithm can be found in the original publication.[15]

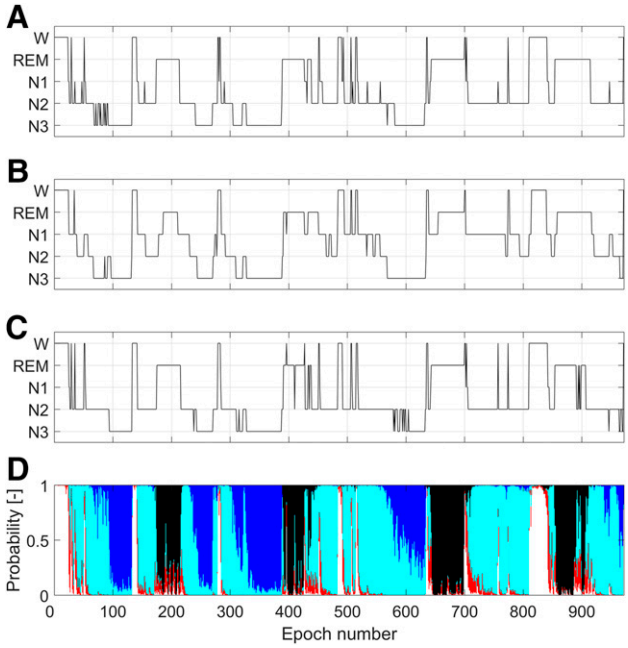
IRR evaluation

To evaluate IRR for sleep stage scoring, we built 5 confusion matrices (CMs) for each participant:

- The first CM was obtained bycomparing the hypnogram manually scored in Innsbruck to the one manually scored in Berlin (INN vs BER), as shown in **Figure 3**. The CM summarizes the relationship between the 2 hypnograms. The diagonal elements correspond to the number of epochs where the scorers were in agreement (ie, number of epoch that both scorers scored as W, N1, N2, N3, and REM sleep). The off-diagonal elements correspond to the number of epochs for which the scorers were not in agreement and indicate also the type of disagreement (eg, in **Figure 3C**, the element in {row 1, column 2} implies that 10 epochs were scored as W by the scorer in Innsbruck, but as N1 by the scorer in Berlin).
- The second CM was derived by comparing the hypnogram obtained from manual scoring in Innsbruck to the one obtained from the automatic scoring (INN vs AUTO). This CM was built similarly to the one in INN vs BER.
- The third CM was made by comparing the manual hypnogram from Berlin to the automatic one (BER vs AUTO). This CM was built similarly to the one in INN vs BER.
- The fourth CM was obtained by comparing the sleep epochs for which manual scorers from Innsbruck and Berlin agreed (ie, consensus epochs) to the respective sleep epochs automatically scored (CONS vs AUTO), as shown in **Figure S1** in the supplemental material.
- The fifth CM was derived by comparing both manual hypnograms to the automatic one (MAN vs AUTO). In case of disagreement between the 2 manual scorings for 1 sleep epoch, an epoch was counted equally in the 2 sleep stages that were manually assigned (**Figure S2** in the supplemental material). This is the same approach presented in the original publication of the Stanford-STAGES algorithm.[15]

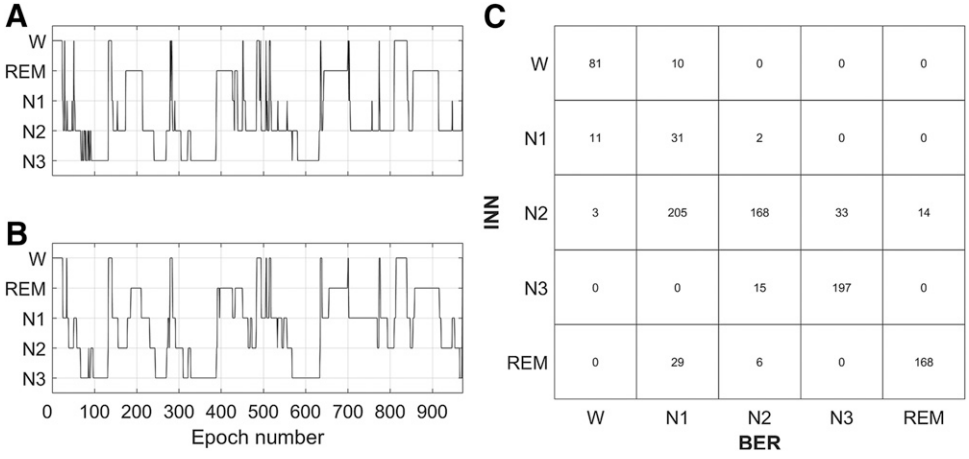
Each CM was used to compute several IRR measures. More specifically, the overall Cohen's κ and the overall accuracy (A;

Figure 2—Visual comparison of hypnograms and hypnodensity for the same polysomnographic recording.



(A) Manual hypnogram scored in Innsbruck. **(B)** Manual hypnogram scored in Berlin. **(C)** Hypnogram obtained by applying the automatic Stanford-STAGES algorithm. For each epoch, the sleep stage assigned was the one having the highest probability in the hypnodensity **(D)**. Color codes for probabilities in the hypnodensity: white, W; red, N1; light blue, N2; dark blue, N3; black, REM. W = wakefulness, REM = rapid eye movement sleep, N1 = non-REM stage 1 sleep, N2 = non-REM stage 2 sleep, N3 = non-REM stage 3 sleep.

Figure 3—Hypnograms and relative confusion matrix.



(Left) Hypnograms for the same PSG recording scored by human scorers in Innsbruck **(A)** and Berlin **(B)** are shown. The confusion matrix **(C)** reports in the diagonal the number of epochs for which the scorers were in agreement and out of the diagonal the number of epochs for which there was disagreement and the type of disagreement (eg, the element in {row 1, column 2} indicates that 10 epochs were scored as W in the INN hypnogram but as N1 in the BER hypnogram). BER = Berlin, INN = Innsbruck, PSG = polysomnography, W = wakefulness, REM = rapid eye movement sleep, N1 = non-REM stage 1 sleep, N2 = non-REM stage 2 sleep, N3 = non-REM stage 3 sleep.

Table 1—Sleep stages distributions in the hypnograms scored in Innsbruck, Berlin, and by the algorithm.

Sleep Stages	Measure	INN	BER	AUTO	*P*		
					INN vs BER	INN vs AUTO	BER vs AUTO
W (% of TIB)	μ ± σ	18.80 ± 12.05	18.95 ± 12.39	24.57 ± 15.01	n.s.	< .001	< .001
	m [5th–95th]	16.49 [4.40–42.49]	16.62 [4.13–42.58]	21.71 [6.61–53.8]			
N1 (% TIB)	μ ± σ	11.44 ± 6.61	20.43 ± 10.12	4.49 ± 3.05	< .001	< .001	< .001
	m [5th–95th]	9.88 [4.48–23.86]	18.55 [7.54–39.83]	3.76 [0.89–10.02]			
N2 (% TIB)	μ ± σ	40.85 ± 9.47	31.41 ± 10.67	47.68 ± 11.44	< .001	< .001	< .001
	m [5th–95th]	41.57 [24.36–55.23]	31.20 [13.65–48.65]	48.59 [27.05–64.20]			
N3 (% TIB)	μ ± σ	10.92 ± 7.01	14.71 ± 7.23	6.10 ± 6.23	< .001	< .001	< .001
	m [5th–95th]	10.85 [0.00–22.99]	14.29 [3.19–26.94]	4.50 [0.00–18.19]			
REM (% TIB)	μ ± σ	14.34 ± 5.78	10.86 ± 5.54	11.91 ± 6.13	< .001	< .001	.001
	m [5th–95th]	14.21 [5.12–23.85]	10.80 [1.45–20.61]	11.90 [2.00–22.43]			

The distributions are calculated as percentage of time in bed (TIB) and shown as mean (μ) ± 1 standard deviation (σ), median (m), and 5th–95th percentiles across the participants. Statistical analyses were performed with Mann-Whitney *U* tests corrected with Bonferroni procedure. Statistical significance was set at the value of .05. AUTO = automatic algorithm, BER = Berlin, INN = Innsbruck, N1 = non-REM stage 1 sleep, N2 = non-REM stage 2 sleep, N3 = non-REM stage 3 sleep, n.s. = nonsignificant; REM = rapid eye movement, W = wakefulness.

ie, the percentage of sleep epochs included in the diagonal of the CM) were calculated from each CM. Stage-specific Cohen's κ (κ_W, κ_{N1}, κ_{N2}, κ_{N3}, κ_{REM}), accuracies (A_W, A_{N1}, A_{N2}, A_{N3}, A_{REM}), and F1 scores ($F1_W$, $F1_{N1}$, $F1_{N2}$, $F1_{N3}$, $F1_{REM}$) were also obtained. Finally, the overall F1 score (F1) was calculated as average across the stage-specific F1 scores. **Figure S3** in the supplemental material provides the equations that were used to compute all the IRR measures. We decided to calculate all these different measures to provide a complete and comprehensive overview of the agreement, as each of them considers different aspects of the agreement.[22] In particular, accuracy gives higher importance to true positives and true negatives, F1 score to false negatives and false positives, and Cohen's κ corrects accuracy for possible random agreement.[22] The values of Cohen's κ were interpreted according to Landis and Koch[23]: $\kappa < 0$ indicating no agreement, $0 \le \kappa < 0.20$ indicating slight agreement, $0.20 \le \kappa < 0.40$ indicating fair agreement, $0.40 \le \kappa < 0.60$ indicating moderate agreement, $0.60 \le \kappa < 0.80$ indicating substantial agreement, and $\kappa \ge 0.80$ indicating almost perfect agreement.

Sleep and demographic factors influencing IRR

The effects of age, sex, body mass index (BMI), PLMS index, and AHI on the overall κ, A, and F1 score values were evaluated by means of multiple linear regression analysis. All predictors (except for sex, which was included as a categorical predictor) and κ, A, and F1 score values were Z-score transformed. Noncollinearity of predictors was ensured by checking that the variation inflation factors were less than 10.[24] Normality of the residuals was checked by visual inspection of Q-Q plots. In case of not normal residuals, we applied cubic transformation to κ, A, and F1 score values.

RESULTS

Out of the 1,066 participants included in the study, 570 were men (53.47%). The median age of the participants was 54 years (5th–95th percentiles: 26–74 years), the median BMI was 28.0 kg/m^2 (5th–95th percentiles: 21.1–37.1 kg/m^2), the median AHI was 3.9 events/h (5th–95th percentiles: 0.1–40.0 events/h), and the median PLMS index was 6.4 (5th–95th percentiles: 0–66.4). The distributions of sleep stages scored by the human scorers and the automatic algorithm across the participants are shown in **Table 1**. No significant difference was found in the percentage of manually scored W, whereas significant differences were present between the percentage of all other sleep stages scored by sleep experts in Innsbruck and Berlin and by the algorithm.

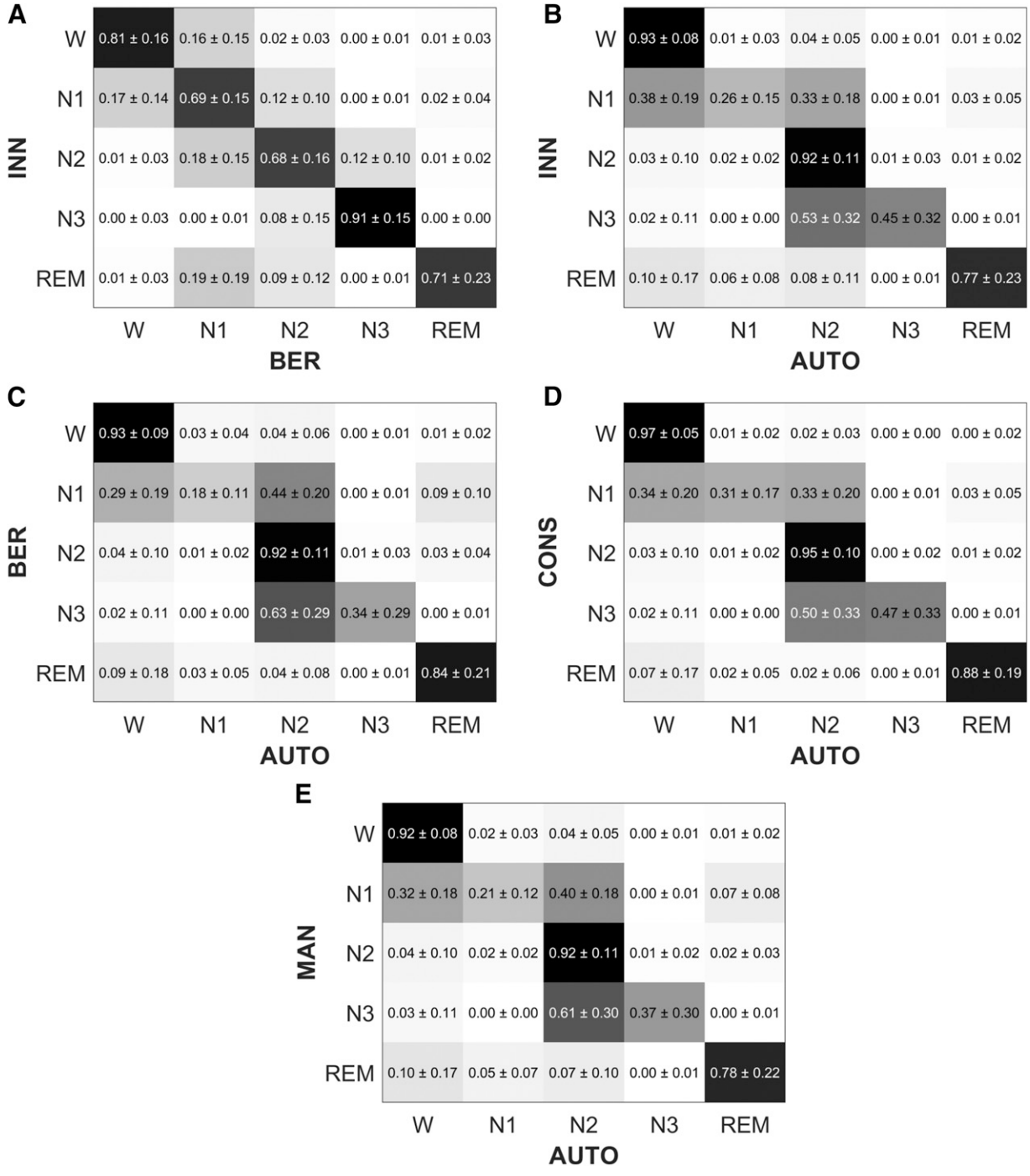
To have an overview of the obtained CMs across the participants, each CM was row-wise normalized (ie, for each row, the elements were divided by their sum). **Figure 4** shows the mean and standard deviation values of the normalized CMs across the participants. As an example to interpret the row-wise normalized CMs, in **Figure 4A**, the element in {row 1, column 1} indicates that 81 ± 16% of the epochs scored as W in Innsbruck were also scored as W in Berlin. Similarly, the element in {row 1, column 2} indicates that 16 ± 15% of the epochs scored as W in Innsbruck were scored as N1 in Berlin.

Table 2 shows the average, standard deviation, median, and 5th and 95th percentiles of the overall κ and stage-specific κ for all comparison across the participants. The values for overall and stage-specific accuracies and F1 scores are reported in **Table S1** and **Table S2**.

The results reported in **Figure 4**, **Table 2**, **Table S1**, and **Table S2** show that there was an average overall substantial agreement between the manual scorers (κ = 0.66 ± 0.13, A = 0.75 ± 0.10, F1 = 0.70 ± 0.10) and the automatic sleep staging algorithm agreed significantly more ($P < .001$ with Mann-Whitney *U* tests for all overall performances) with the scorings performed in Innsbruck (κ = 0.68 ± 0.14, A = 0.78 ± 0.10, F1 = 0.65 ± 0.13) than to the ones performed in Berlin (κ = 0.55 ± 0.15, A = 0.67 ± 0.12, F1 = 0.59 ± 0.12).

When the automatic scoring was compared with the epochs where there was an agreement between manual scorers (ie,

Figure 4—Row-wise normalized confusion matrices across all participants.



The values are shown as mean and standard deviation across the participants. For each matrix element, a darker color represents a higher agreement. **(A)** INN vs BER: comparison of manual hypnograms scored in Innsbruck and Berlin. **(B)** INN vs AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones. **(C)** BER vs AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones. **(D)** CONS vs AUTO: comparison of the epochs where manual scorers from Innsbruck and Berlin were in consensus to the respective epochs automatically scored. **(E)** MAN vs AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the 2 manually scored stages). As an example to interpret these row-wise CMs, in **A**, the element in {row 1, column 1} indicates that 81 ± 16% of the epochs scored as W in Innsbruck were also scored as W in Berlin. Similarly, the element in {row 1, column 2} indicates that 16 ± 15% of the epochs scored as W in Innsbruck were scored as N1 in Berlin. W = wakefulness, REM = rapid eye movement sleep, N1 = non-REM stage 1 sleep, N2 = non-REM stage 2 sleep, N3 = non-REM stage 3 sleep.

CONS vs AUTO), substantial values of overall agreement were obtained ($\kappa = 0.73 \pm 0.14$, A = 0.82 ± 0.10, F1 = 0.70 ± 0.13), indicating that the automatic scoring overall correctly classified the sleep epochs with clearer patterns. When both manual scorings were compared with the automatic ones (ie, MAN vs AUTO), the overall average performances ($\kappa = 0.61 \pm 0.14$, A =

Table 2—Overall and stage-specific values of Cohen's κ for the different comparisons.

Parameter	Measures	Comparison				
		INN vs BER	INN vs AUTO	BER vs AUTO	CONS vs AUTO	MAN vs AUTO
κ	$\mu \pm \sigma$	0.66 ± 0.13	0.68 ± 0.14	0.55 ± 0.15	0.73 ± 0.14	0.61 ± 0.14
	m [5th–95th]	0.67 [0.43–0.83]	0.70 [0.41–0.84]	0.56 [0.30–0.77]	0.76 [0.47–0.90]	0.62 [0.36–0.79]
κ_W	$\mu \pm \sigma$	0.78 ± 0.14	0.74 ± 0.17	0.70 ± 0.19	0.80 ± 0.17	0.72 ± 0.17
	m [5th–95th]	0.81 [0.48–0.94]	0.78 [0.38–0.92]	0.75 [0.31–0.92]	0.86 [0.43–0.96]	0.76 [0.39–0.91]
κ_{N1}	$\mu \pm \sigma$	0.40 ± 0.16	0.30 ± 0.16	0.22 ± 0.14	0.39 ± 0.20	0.25 ± 0.14
	m [5th–95th]	0.40 [0.13–0.66]	0.30 [0.04–0.57]	0.21 [0.02–0.47]	0.40 [0.05–0.70]	0.24 [0.04–0.48]
κ_{N2}	$\mu \pm \sigma$	0.62 ± 0.16	0.71 ± 0.14	0.53 ± 0.17	0.74 ± 0.15	0.62 ± 0.14
	m [5th–95th]	0.65 [0.34–0.84]	0.73 [0.46–0.87]	0.55 [0.24–0.79]	0.77 [0.46–0.92]	0.63 [0.37–0.81]
κ_{N3}	$\mu \pm \sigma$	0.66 ± 0.28	0.49 ± 0.33	0.40 ± 0.31	0.53 ± 0.34	0.42 ± 0.32
	m [5th–95th]	0.77 [0.00–0.93]	0.57 [0.00–0.91]	0.41 [0.00–0.88]	0.64 [0.00–0.94]	0.46 [0.00–0.87]
κ_{REM}	$\mu \pm \sigma$	0.75 ± 0.21	0.79 ± 0.21	0.74 ± 0.23	0.86 ± 0.21	0.76 ± 0.20
	m [5th–95th]	0.82 [0.26–0.96]	0.86 [0.30–0.96]	0.81 [0.15–0.95]	0.94 [0.36–0.99]	0.82 [0.30–0.95]

The values are shown as mean (μ) ± 1 standard deviation (σ), median (m), and 5th– 95th percentiles across the participants. INN vs BER: comparison of manual hypnograms scored in Innsbruck and Berlin; INN vs AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones; BER vs AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones; CONS vs AUTO: comparison of the epochs where manual scorers from Innsbruck and Berlin were in consensus to the respective epochs automatically scored; MAN vs AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the two manually scored stages). AUTO = automatic algorithm, BER = Berlin, CONS = consensus, INN = Innsbruck, MAN = manual, N1 = non-REM stage 1 sleep, N2 = non-REM stage 2 sleep, N3 = non-REM stage 3 sleep, REM = rapid eye movement, W = wakefulness.

0.72 ± 0.01, F1 = 0.62 ± 0.12) ranged between to the ones obtained for INN vs AUTO and BER vs AUTO.

Concerning the single sleep stage performances, the accuracy values (**Table S2**) were always high because of imbalanced classes; therefore, a meaningful analysis of the agreements can be performed only considering κ and F1 score values. The manual scorers from Innsbruck and Berlin tended to disagree mostly for scoring of N1 sleep (κ_{N1} = 0.40 ± 0.16, $F1_{N1}$ = 0.49 ± 0.15), whereas the agreement was substantial for all the other sleep stages.

Compared with the 2 manual scorings, the automatic scoring had an average fair agreement for N1 (κ_{N1} = 0.30 ± 0.16, $F1_{N1}$ = 0.34 ± 0.16 for INN vs AUTO and κ_{N1} = 0.22 ± 0.14, $F1_{N1}$ = 0.28 ± 0.15 for BER vs AUTO). For N3 sleep, the average agreement between the human scorers and the algorithm was moderate (κ_{N3} = 0.49 ± 0.33, $F1_{N3}$ = 0.48 ± 0.35 for INN vs AUTO and κ_{N3} = 0.40 ± 0.31, F_{N3} = 0.43 ± 0.33 for BER vs AUTO). The agreement was, on average, substantial for all the other sleep stages, except for N2 sleep for BER vs AUTO (moderate agreement).

The stage-specific performances for the comparison MAN vs AUTO ranged between the ones obtained for INN vs AUTO and BER vs AUTO. Concerning the comparison CONS vs AUTO, there was an average substantial agreement for W, N2 and REM sleep (κ_W = 0.80 ± 0.17, $F1_W$ = 0.85 ± 0.14, κ_{N2} = 0.74 ± 0.15, $F1_{N2}$ = 0.84 ± 0.11, κ_{REM} = 0.86 ± 0.21, $F1_{REM}$ = 0.86 ± 0.23), whereas only a fair average agreement was obtained for N1 sleep (κ_{N1} = 0.39 ± 0.20, $F1_{N1}$ = 0.42 ± 0.19) and a moderate one for N3 sleep (κ_{N3} = 0.53 ± 0.34, $F1_{N3}$ = 0.50 ± 0.37).

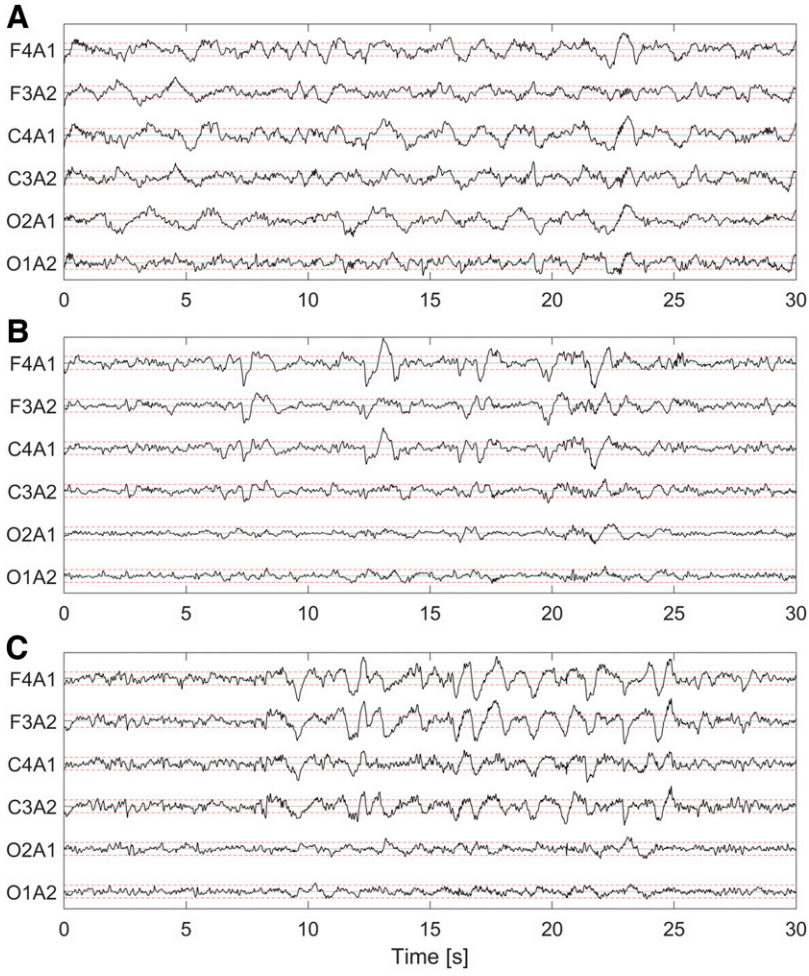
It can be noticed that the performances obtained for N3 sleep have median values of κ_{N3} and $F1_{N3}$ for CONS vs AUTO around 0.10 higher than the average ones. This is because the automatic algorithm did not score any epoch as N3 sleep for 193 participants, whereas no N3 sleep was scored for 105 participants in Innsbruck and for 5 participants in Berlin. In general, we observed that the automatic algorithm scored significantly less N3 sleep compared with the human scorers (**Table 1**). To better understand the possible cause of disagreement between manual and automatic scoring, **Figure 5** shows (A) an epoch in which the scorer correctly scored N3 sleep (as 24% of the epoch contains slow wave activity) but the algorithm scored N2 sleep; (B) an epoch in which the human scorers wrongly scored N3 sleep (as only 17% of the epoch contains slow wave activity), but the algorithm correctly scored N2 sleep; and (C) an epoch where human scorers and the algorithm correctly scored N3 sleep (slow wave activity present in 42% of the epoch). The percentages of slow wave activity in these epochs were manually counted by an independent scorer.

Table 3 shows the results of the multiple regression analyses for the overall κ values, and **Table S3** and **Table S4** show the ones for accuracy and F1 score values. A higher disagreement for sleep stage scoring was observed for men and older participants with higher AHI and BMI.

DISCUSSION

We evaluated IRR for sleep stage scoring performed by human scorers in 2 European sleep centers and by the artificial intelligence–based Stanford-STAGES algorithm on a population-based study in 1,066 participants, who were part of the Study of Health in Pomerania-TREND baseline examination. The results showed an overall substantial agreement for sleep stage scoring between the 2 sleep centers and between human and automatic scoring. The human scorers disagreed mostly for N1 sleep and the automatic scoring had the lowest agreement with the manual scoring for N1 and N3 sleep. Furthermore, we assessed that sleep stage scoring

Figure 5—Example of sleep epochs where manual scorers agreed to score N3 sleep.



(A) Epoch that was correctly manually scored in both centers as N3 sleep (slow wave activity covers 24% of the epoch), but was scored as N2 by the algorithm. **(B)** Epoch that was wrongly scored as N3 by human scorers in the 2 sleep centers (17% of the epoch contains slow wave activity) but correctly scored as N2 by the algorithm. **(C)** Epoch correctly scored by the algorithm and the human scorers as N3 sleep (42% of the epoch has slow wave activity). For each electroencephalographic channel, the red lines are drawn at −37.5 and +37.5 μV to highlight the amplitude of 75-μV peak-to-peak amplitude of slow waves. N2 = non–rapid eye movement (NREM) stage 2 sleep, N3 = NREM stage 3 sleep.

disagreement is higher for older participants, for men, and for participants with higher AHI and BMI.

The overall agreement between scorers in Innsbruck and Berlin is in the range reported in previous studies.[2–7] There was substantial agreement for all sleep stages, except for N1 sleep. The lower agreement for scoring N1 sleep matches previous literature results[3–5,7] and substantiates the difficulties in scoring this stage, because of its transitory nature between wakefulness and deeper sleep[1] and to the lack of specific electrophysiologic elements (eg, as sleep spindles, K-complexes, delta waves, or rapid eye movements). All the stage-specific κ values lay in the range of the values reported in previous studies.[3–5,7] Therefore, our findings confirm the rates of IRR for sleep stage scoring between human scorers previously presented in literature.[3–5,7]

The automatic scoring agreed more with the manual scoring performed in Innsbruck than the one carried out in Berlin. Sleep recordings from Innsbruck were included in the original work where the Stanford-STAGES algorithm was presented.[15] However, they were not used to train the algorithm, thus ensuring no bias in our results. As the algorithm mostly disagreed with the human scorers for N1 and N3 sleep, a possible reason for the lower overall agreement for BER vs AUTO compared with INN vs AUTO might be the higher number of epochs scored as N1 and N3 sleep in Berlin.

The Stanford-STAGES algorithm has been originally tested on the Inter-Scorer Reliability Cohort, a cohort of 70 PSGs scored by 6 sleep experts,[25] where it achieved an average accuracy of 0.78 compared with manual sleep scoring.[15] Because the Inter-Scorer Reliability Cohort includes only women[25] and male sex significantly decreases IRR, such accuracy should be compared with the average one we obtained when considering only women from our cohort (ie, 0.75 for MAN vs AUTO). This comparison shows that, in our cohort, the algorithm performed only slightly worse than in the Inter-Scorer Reliability Cohort. In the original study, the Stanford-STAGES algorithm has also been tested in 3 other cohorts,[26–29] all scored only by 1 single

Table 3—Results of the multiple regression linear analyses for overall Cohen's κ.

Predictors	INN vs BER		INN vs AUTO*		BER vs AUTO		CONS vs AUTO*		MAN vs AUTO*	
	b	*P*	*b*	*P*	*b*	*P*	*b*	*P*	*b*	*P*
Intercept	−0.118	.003	−0.145	< .001	−0.188	< .001	−0.192	< .001	−0.208	< .001
Age	−0.106	< .001	−0.060	.053	−0.145	< .001	−0.097	.002	−0.132	< .001
Sex (F)	0.251	< .001	0.308	< .001	0.399	< .001	0.407	< .001	0.441	< .001
PLMS index	−0.023	.443	−0.047	.114	−0.033	.272	−0.057	.054	−0.042	.152
AHI	−0.227	< .001	−0.238	< .001	−0.175	<.001	−0.185	< .001	−0.200	< .001
BMI	−0.041	.189	−0.087	.005	−0.066	.031	−0.077	.012	−0.080	.008
Overall *P*	< .001		< .001		< .001		< .001		< .001	

For each analysis, the overall Cohen's κ was the outcome variable and age, sex (categorical), PLMS index, AHI, and BMI the predictors. Z-score transformations were applied to both outcome variable and predictors (except sex). For each model, the overall *P* value is reported, as well as the slope estimate (*b*) and the *P* value of each predictor. Statistical significance was set at the value of .05. INN vs BER: comparison of manual hypnograms scored in Innsbruck and Berlin; INN vs AUTO: comparison of manual hypnograms scored in Innsbruck to the automatic ones; BER vs AUTO: comparison of manual hypnograms scored in Berlin to the automatic ones; CONS vs AUTO: comparison of the epochs where manual scorers from Innsbruck and Berlin were in consensus to the respective epochs automatically scored; MAN vs AUTO: comparison of both manual hypnograms to the automatic one (in case of disagreement between manual scorers, an epoch was equally weighted between the 2 manually scored stages). AHI = apnea-hypopnea index, AUTO = automatic algorithm, BER = Berlin, BMI = body mass index, CONS = consensus, INN = Innsbruck. MAN = manual, PLMS = periodic limb movement during sleep. *Cubic transformation was applied to Cohen's κ in the highlighted comparisons to meet the normality assumption of the model residuals.

scorer, achieving average accuracy between 0.77 and 0.86.[15] In our cohort, the automatic scoring had an average overall accuracy between 0.67 (for BER vs AUTO) and 0.78 (for INN vs AUTO). Therefore, when considering the performances obtained for INN vs AUTO, the Stanford-STAGES algorithm showed similar performances to the ones reported in the original study. This suggests an overall good generalizability of the algorithm to a new unseen cohort.

The statistical analyses revealed that increased age, male sex, and higher AHI and BMI are related to higher disagreement in sleep stage scoring. This is likely because of the increased sleep fragmentation and sleep structure changes seen in the elderly,[30,31] the lower sleep quality in men compared with women,[32] the increased sleep fragmentation caused by increased end-apneic arousals,[33] and the increased sleep disruption associated with higher BMI.[34]

Concerning single sleep stage performances, the biggest discrepancies compared with the manual hypnograms were found in the automatic scoring of N1 and N3 sleep. Previous studies have reported agreements (measured by κ) in the range of 0.16–0.57 for N1 sleep[3–5,7] between human scorers. The agreement values we found between automatic and manual scoring lay in this range.

For N3 sleep, previous studies have reported agreements (measured by κ) between human scorers in the range of 0.49–0.79.[3–5,7] On average, we found that the agreement between the Stanford STAGES-algorithm and the human scorers in Berlin and Innsbruck was lower, because the algorithm tended to score less epochs as N3 sleep. This could be because of several reasons. First, the automatic algorithm used only central and occipital EEG derivations, whereas the human scorers evaluated also the frontal EEG electrodes while scoring. As the slow waves characterizing N3 sleep are most prominent in the frontal derivations[1] and as AASM recommends using the frontal derivations to measure slow wave amplitudes,[19,20] the human scorers might have been more sensitive for recognizing slow wave activity (**Figure 5A**). Second, the underscoring of N3 sleep by the algorithm could be the consequence of the training process of the algorithm, where very few epochs scored as N3 sleep might have been included and no strategy to overcome class imbalance might have been used.[15] Third, this discrepancy might also be caused by the human scorers being too sensitive to visual identification of slow wave activity (**Figure 5B**). Previous findings showed that some human raters tend to score N3 even in presence less than 20% of the epoch covered by slow wave activity.[35] An in-depth analysis of the causes of the discrepancies for scoring N3 sleep is not the main aim of this work. However, the results indicate possible issues of the Stanford-STAGE algorithm in scoring N3 sleep. Future independent studies should further investigate this.

Improvements for the generalizability of the Stanford-STAGES algorithm could be achieved by applying transfer learning, a technique that allows adaptation of a pretrained algorithm to unseen data to improve the classification performances. Some studies have already shown promising results in adapting a pretrained artificial intelligence algorithm for sleep stage scoring in new cohorts.[36–38] However, current transfer learning techniques still require expert knowledge in manual fine-tuning of the algorithms, making their implementation in a clinical environment impractical. Future studies should evaluate automatic procedures to apply transfer learning in the context of automatic sleep stage scoring.

Our results suggest that the Stanford-STAGES algorithm is overall generalizable to a new cohort and therefore potentially applicable in clinical practice; further validations in different cohorts are needed before its integration in clinical routine. In a first step, the algorithm could be used for semiautomatic sleep stage scoring. In particular, the hypnodensity could be used as a useful tool for this purpose, as a sleep expert could perform just a fast check for the epochs where a stage has a clearly higher

probability than the other stages, whereas a more careful check would be required only for the remaining epochs. Previous studies showed that manual edit of an automatic sleep stage scoring substantially reduces the interrater variability[6,8,39] and that a semiautomatic approach is less time consuming than a complete manual analysis.[40] Future research should investigate the usefulness of the Stanford-STAGES algorithm in the context of semiautomated sleep stage scoring.

The main limitation of this work is that our study population belongs to a population-based cohort and thus does not represent the population usually admitted in a sleep center. Therefore, our results might be not representative of the performances of sleep stage scoring in a clinical environment, where current scoring rules are particularly challenging to apply to patients (eg, with neurodegeneration).[41] Furthermore, as previously outlined, 2 different versions of the AASM manual for scoring sleep were used in the 2 different European centers[19,20] (these criteria slightly differ for the definitions of transitions N1-N2 and for the definition of REM sleep[42]), thus the results might have a bias. Another potential limitation is that sleep stages were scored in Berlin by different sleep experts, whereas in Innsbruck it was only by one expert. Despite interrater variability between scorers in Berlin was minimized,[18] the inclusion of different experts in Berlin might constitute a bias for the evaluation of IRR, which could not be considered in our analysis. Furthermore, different years of experience in sleep stage scoring of the different experts might have influenced our results. Finally, when we compared the performances achieved in this cohort to the ones obtained in the cohorts where the Stanford-STAGES algorithm was originally validated, we could not take into account possible difference in age, sex distribution, AHI, and BMI between the cohorts. These factors might be the cause of the slightly lower agreements seen in our cohort.

In conclusion, our study evaluated IRR for sleep stage scoring between 2 European sleep centers and the automatic artificial intelligence–based Stanford-STAGES algorithm. Our results show that IRR between human scorers is similar to previously reported results. Furthermore, we found that the Stanford-STAGES algorithm had an overall good agreement with manual scoring and that it performed similarly in this cohort as the in ones where it has been originally tested, thus suggesting that it is generalizable to new unseen cohorts. Future studies should further confirm our findings in new independent cohorts. Future research should also evaluate the integration of automatized transfer learning techniques for better adaptation of the Stanford-STAGES algorithm to new cohorts and the integration of the algorithm for semiautomated sleep stage scoring for clinical purposes.

ABBREVIATIONS

A, accuracy
AASM, American Academy of Sleep Medicine
AHI, apnea-hypopnea index
AUTO, Automatic
BER, Berlin
BMI, body mass index
CM, confusion matrix
CONS, Consensus
EEG, electroencephalogram
EDF, European data format
INN, Innsbruck
IRR, interrater reliability
MAN, Manual
N1, non–rapid eye movement sleep stage 1
N2, non–rapid eye movement sleep stage 2
N3, non–rapid eye movement sleep stage 3
PLMS, periodic limb movements during sleep
PSG, polysomnography
REM, rapid eye movement
W, wakefulness

REFERENCES

1. Berry RB, Quan SF, Abreu AR, et al; for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Version 2.6. Darien, IL: American Academy of Sleep Medicine; 2020.
2. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med*. 2013;9(1):81–87.
3. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *J Sleep Res*. 2009;18(1):74–84.
4. Zhang X, Dong X, Kantelhardt JW, et al. Process and outcome for international reliability in sleep scoring. *Sleep Breath*. 2015;19(1):191–195.
5. Deng S, Zhang X, Zhang Y, et al. Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard. *Sleep Breath*. 2019;23(2):719–728.
6. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. *J Clin Sleep Med*. 2016;12(6):885–894.
7. Magalang UJ, Chen N-H, Cistulli PA, et al; SAGIC Investigators. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep*. 2013;36(4):591–596.
8. Younes M, Hanly PJ. Minimizing interrater variability in staging sleep by use of computer-derived features. *J Clin Sleep Med*. 2016;12(10):1347–1356.
9. Itil TM. Automatic classification of sleep stages and the discrimination of vigilance changes using digital computer methods. *Agressologie*. 1969;10(suppl):603–610.
10. Penzel T, Hirshkowitz M, Harsh J, et al. Digital analysis and technical specifications. *J Clin Sleep Med*. 2007;3(2):109–120.
11. Boostani R, Karimzadeh F, Nami M. A comparative review on sleep stage classification methods in patients and healthy individuals. *Comput Methods Programs Biomed*. 2017;140:77–91.
12. Lajnef T, Chaibi S, Ruby P, et al. Learning machines and sleeping brains: Automatic sleep stage classification using decision-tree multi-class support vector machines. *J Neurosci Methods*. 2015;250:94–105.
13. Fiorillo L, Puiatti A, Papandrea M, et al. Automated sleep scoring: a review of the latest approaches. *Sleep Med Rev*. 2019;48:101204.
14. Goldstein CA, Berry RB, Kent DT, et al. Artificial intelligence in sleep medicine: background and implications for clinicians. *J Clin Sleep Med*. 2020;16(4):609–618.
15. Stephansen JB, Olesen AN, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat Commun*. 2018;9(1):5229.
16. Lim DC, Mazzotti DR, Sutherland K, et al. Reinventing polysomnography in the age of precision medicine. *Sleep Med Rev*. 2020;52:52.
17. Völzke H, Alte D, Schmidt CO, et al. Cohort profile: the study of health in Pomerania. *Int J Epidemiol*. 2011;40(2):294–307.

18. Stubbe B, Penzel T, Fietze I, et al. Polysomnography in a large population based study: the Study of Health in Pomerania protocol. *J Sleep Disord Manag*. 2016;2:10.
19. Iber C, Ancoli-Israel S, Chesson AL, Quan SF; for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. 1st ed. Westchester, IL: American Academy of Sleep Medicine; 2007.
20. Berry RB, Brooks R, Gamaldo CE, et al; for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Version 2.0. Darien, IL: American Academy of Sleep Medicine; 2012.
21. Szentkirályi A, Stefani A, Hackner H, et al. Prevalence and associated risk factors of periodic limb movement in sleep in two German population-based studies. *Sleep*. 2019;42(3):zsy237.
22. Ballabio D, Grisoni F, Todeschini R. Multivariate comparison of classification performance measures. *Chemom Intell Lab Syst*. 2018;174:33–44.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
24. Kutner M, Nachtsheim C, Neter J. *Applied Linear Regression Models*. 4th ed. New York: McGraw-Hill Irwin; 2004.
25. Kuna ST, Benca R, Kushida CA, et al. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *Sleep*. 2013;36(4):583–589.
26. Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. *N Engl J Med*. 1993; 328(17):1230–1235.
27. Andlauer O, Moore H, Jouhier L, et al. Nocturnal rapid eye movement sleep latency for identifying patients with narcolepsy/hypocretin deficiency. *JAMA Neurol*. 2013;70(7):891–902.
28. Moore H, Leary E, Lee S-Y, et al. Design and validation of a periodic leg movement detector. *PLoS One*. 2014;9:e114565.
29. Hong SC, Lin L, Jeong JH, et al. A study of the diagnostic utility of HLA typing, CSF hypocretin-1 measurements, and MSLT testing for the diagnosis of narcolepsy in 163 Korean patients with unexplained excessive daytime sleepiness. *Sleep*. 2006;29(11):1429–1438.
30. Mander BA, Winer JR, Walker MP. Sleep and human aging. *Neuron*. 2017; 94(1):19–36.
31. Cesari M, Stefani A, Mitterling T, Frauscher B, Schönwald SV, Högl B. Sleep modelled as a continuous and dynamic process predicts healthy ageing better than traditional sleep scoring. *Sleep Med*. 2021;77:136–146.
32. Goel N, Kim H, Lao RP. Gender differences in polysomnographic sleep in young healthy sleepers. *Chronobiol Int*. 2005;22(5):905–915.
33. Kimoff RJ. Sleep fragmentation in obstructive sleep apnea. *Sleep*. 1996;19(9 Suppl):S61–S66.
34. van den Berg JF, Knvistingh Neven A, Tulen JHM, et al. Actigraphic sleep duration and fragmentation are related to obesity in the elderly: the Rotterdam Study. *Int J Obes Lond*. 2008;32(7):1083–1090.
35. Younes M, Kuna ST, Pack AI, et al. Reliability of the American Academy of Sleep Medicine rules for assessing sleep depth in clinical practice. *J Clin Sleep Med*. 2018;14(2):205–213.
36 Phan H, Chén OY, Koch P, Mertins A, De Vos M. Deep transfer learning for single-channel automatic sleep staging with channel mismatch. Paper presented at: 2019 27th European Signal Processing Conference (EUSIPCO); September 2–6, 2019; A Coruna, Spain, 2019.
37. Phan H, Chén OY, Koch P, et al. Towards more accurate automatic sleep staging via deep transfer learning. *IEEE Trans Biomed Eng*. 2020. In press.
38. Abou Jaoude M, Sun H, Pellerin KR, et al. Expert-level automated sleep staging of long-term scalp EEG recordings using deep learning. *Sleep*. 2020;43(11): zsaa112.
39. Svetnik V, Ma J, Soper KA, et al. Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. *Sleep*. 2007;30(11):1562–1574.
40. Younes M, Thompson W, Leslie C, Egan T, Giannouli E. Utility of technologist editing of polysomnography scoring performed by a validated automatic system. *Ann Am Thorac Soc*. 2015;12(8):1206–1218.
41. Santamaria J, Högl B, Trenkwalder C, Bliwise D. Scoring sleep in neurological patients: the need for specific considerations. *Sleep*. 2011;34(10): 1283–1284.
42. American Academy of Sleep Medicine. The 2007 AASM Scoring Manual vs the AASM Scoring Manual v2.0. https://j2vjt3dnbra3ps7ll1clb4q2-wpengine.netdna-ssl.com/wp-content/uploads/2017/11/Summary-of-Updates-in-v2.0-FINAL.pdf. Accessed September 1, 2020.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication October 1, 2020
Submitted in final revised form January 25, 2021
Accepted for publication January 25, 2021
Address correspondence to: Matteo Cesari, PhD, Medical University of Innsbruck, Department of Neurology, Anichstrasse 35, 6020 Innsbruck, Austria;
Email: matteo.cesari@i-med.ac.at

DISCLOSURE STATEMENT

All authors have seen this manuscript and approved its submission. Work for this study was performed at the Department of Neurology, Medical University of Innsbruck. Study of Health in Pomerania is part of the Community Medicine Research Network of the University Medicine Greifswald, which is supported by the German Federal State of Mecklenburg-West Pomerania. Polysomnography assessment was in part supported by the German RLS organization (Deutsche Restless Legs Vereinigung). The authors report no conflicts of interest.